# Bridging the Gaps between Digital Humanities, Lexicography, and Linguistics: A TEI Dictionary for the Documentation of Mixtepec-Mixtec

Jack Bowers, Laurent Romary

## Reference Works in Progress

# Bridging the Gaps between Digital Humanities, Lexicography, and Linguistics: A TEI Dictionary for the Documentation of Mixtepec-Mixtec

Jack Bowers and Laurent Romary
jack.bowers@oeaw.ac.at laurent.romary@inria.fr

**ABSTRACT**

This paper discusses the digital dictionary component in an ongoing language documentation project for the Mixtepec-Mixtec language (iso 639-3: mix). Mixtepec-Mixtec (Sa'an Savi 'rain language') is an Otomonguean language spoken by roughly 9,000–10,000 people in the Juxtlahuaca district of Oaxaca and in parts of the Guerrero and Puebla states of Mexico. Creating a digital dictionary for an under-resourced language entails a number of challenges that require unique and nuanced encoding solutions in which a delicate balance between the linguistic content, data structure, potential linked resources, and editorial metadata must be found. Herein we demonstrate how we use TEI to create a reusable, extensible, and machine readable language resource with an emphasis on how our solutions using a combination of novel and established TEI dictionary structures enable us to address our specific needs for Mixtepec-Mixtec and also provide a relevant roadmap for similar under-resourced language projects.

**Keywords**: language documentation, digital humanities, Mixtec, TEI, dictionary encoding

## INTRODUCTION

This paper discusses the dictionary component of a larger documentation project of the Mixtepec-Mixtec language (MIX)[1] using the Text Encoding Initiative, or TEI (www.tei-c.org). In addition to the creation of a TEI dictionary, the primary output of the project is an open source body of reusable and extensible multimedia language resources including a text corpus of spoken and written language encoded and annotated in TEI. The language resources created are in turn being used to further knowledge of all aspects of the language itself within the fields of linguistics and lexicography by producing empirical corpus-based descriptions and analyses of aspects of the language's features.

In the process of data collection, annotation, and encoding, we seek to capture content relevant to every linguistic level from phonetic to semantic and etymological, as well as potential sub-dialectal variation. In conjunction with the complexity of the data, given the maximally broad scope of linguistic and lexicographic research being pursued, it is essential to have a means of organizing all the various components of the languages resources (LR) within a dynamic and flexible system. Also, given the lack of any other dictionary resources for the language, it is especially important that what we create is reusable and extensible so that it may continue to evolve, with the possibility of being easily exported or converted to other formats.

In pursuit of these goals, TEI essentially meets all the needs as the primary format for encoding and annotating our corpus, our born-digital dictionary, and our metadata. TEI is widely accepted in the digital lexicographic community as the de facto standard for the encoding of both retro-digitized and born-digital dictionaries and is being increasingly used for annotated lexical text corpora.

While TEI is well established and increasingly more widely adopted for projects and resources dealing with major world languages,

[1] Most of the spoken data collected in this project originate from consultation sessions with two native speakers from Yucunani (17.30083, -97.89389), a town that is part of the municipality of San Juan de Mixtepec (http://www.geonames.org/3518634/san-juan-mixtepec.html).

particularly those of Europe and North America, it is far less adopted in projects dealing with indigenous languages. Aside from publications related to the current project, Czaykowska-Higgins et al. (2014), describing an application of TEI to the indigenous language Moses-Columbia Salish "Nxaʔamxcín", is a noteworthy example.

Our use of TEI for documentation work requires us to make use of the vocabulary for new or less common applications in order to accommodate the particular nuances of our data. Doing this benefits not only our project: in mapping out how to accommodate new unique combinations of features for a non-Indo-European indigenous language, it also increases the usability of TEI for potential future users and projects seeking to do similar things.

**General issues in dealing with an under-resourced language.** While, according to Ethnologue (Simons and Fennig 2018), the status of MIX is "vigorous," it is an under-resourced language, and that gives rise to several significant challenges in our work, including these:

1. MIX lacks established linguistic descriptions beyond those of phonology and morphology upon which to build;
2. given the limited sample size of our data, and the disproportionate amount of data from two primary consultants from a small village outside the main town of San Juan Mixtepec, it is unclear whether certain observations of variation are due to sub-dialectal differences or speaker demographic-specific factors;
3. because speakers attempting to write in MIX are often not aware of phonological minimal distinctions, further consultation is almost always needed when interpreting and annotate such written materials;
4. because the orthography is still under development[2], the work done in creating and editing orthographic data in the project remains subject to revision;
5. the combination of the issues above and the fact that the output of this data will be the first (at least publicly available) set of MIX

[2] The MIX orthography is under development by Mille Nieves and Gisela Beckmann of SIL Mexico in consultation with native speakers.

annotated data means there will be no way to train a learning model[3] to automate the corpus annotation work.[4]

**Data: sources, formatting, tools, and markup.** Most of the MIX language resources have been collected from recordings of consultation sessions with native speakers[5] and a collection of children's text booklets published by the Summer Institute of Linguistics (SIL) Mexico.[6] Other primary sources of data are written material created by native speakers as part of this project; documents about Mixtec that contain examples from others researchers;[7] a set of public safety documents published by the Mexican government; excerpts from written personal communication with speakers; a small number of previous academic publications on the language (e.g., Pike and Ibach 1978; Paster and Beam de Azcona 2004a, 2004b; Paster 2005, 2010),[8] as well as recordings and videos found online. For our editing and management of the text data we use Oxygen XML editor, and we backup and store our files (excluding soundfiles) on GitHub.[9] All of the project files (including sound and video) are archived in the Harvard Dataverse repository[10] and will be registered with the Open Language Archives (OLAC).

**Fitting in with the standards.** In addition to using TEI for an indigenous language dictionary, Czaykowska-Higgins et al. (2014) cover

[3] An exception to this might be in automatic annotation of phonetic units, but this would not apply to tones because in the data contained so far there is a large degree of ambiguity in tonal classification, owing in part to a lack of diversity in speakers and, in some cases, poor recording quality.

[4] This issue is significant because MIX is tonal and the orthography in previous SIL Mexico publications in the language (which make up most of the text data sources of this project) did not represent tone, resulting in a large number of homographs.

[5] Recorded speech data are transcribed using Praat (Boersma and Weenik 2017) and then converted to TEI-XML using XSLT, where the text transcription output is integrated into the corpus. In accordance with the recommendations of Austin (2006), original recordings are stored in uncompressed 44kHz wav files.

[6] http://www.mexico.sil.org/es/lengua_cultura/mixteca/mixteco-mixtepec-mix [7] Such sources are unpublished notes, obtained from personal communication with Mille Nieves of SIL Mexico.

[8] The primary speaker consultant for the Pastor and Beam de Azcona papers was Juan "Tisu'ma" Salazar, the same primary speaker consultant for the current project.

[9] https://github.com/iljackb/Mixtepec_Mixtec

[10] As of the time of submission, the repository is still in preparation. When published it will be available at https://doi.org/10.7910/DVN/BF2VNK

many other issues relevant to the current project: in particular, questions of standardized lexical terminology, and compatibility with other formats and standards.

In the current project, the issue of deciding upon a theoretically sound fixed lexical terminology is challenging because, given the lack of published linguistic descriptions, there is very little to compare it to, and (as with much else in linguistics) there are multiple overlapping terms to describe the same phenomena.[11] On the technical/standards side, these issues are further complicated by the in-flux state of the standardized repositories created to deal with this very question, namely ISOcat and the CLARIN Concept Registry.[12] Additionally, there are still numerous linguistic concepts that are a significant part of our description (in both the corpus and the dictionary) for which there is no entry in any of the registries;[13] thus, when the given registries are finally stable, as part of this project, we are tasked with submitting the necessary proposals to add entries for certain needed linguistic features.

## TEI DICTIONARY

In this section we give an overview of the components of our dictionary and explain the lexicographic, functional, and—in certain cases—theoretical basis of their use.

Aside from the media files and annotated corpus, the main output of the project will be a trilingual TEI dictionary (Mixtepec-Mixtec, English, and Spanish) containing entries for all glossed lexical items observed in

[11] The SIL-Mexico researchers in Oaxaca have provided a partial inventory of their working terminology; other than that, the only available publications are linguistic descriptions of related Mixtec varieties, which is limited in its utility in certain cases owing to theoretical differences.

[12] See Ide and Romary (2004) for an initial discussion of this that has led to the (now stalled) ISOcat registry. The data corresponding to the latest active state of ISOcat are still statically available from https://old.datahub.io/dataset/isocat in various formats. The CLARIN Concept Registry (see https://www.clarin.eu/ccr) has taken up some of these to provide them as linked open data, but the status of many of these concepts is not yet finalized.

[13] This is also an issue with GOLD (General Ontology for Linguistic Description [Farrar and Langendoen 2003]), a high quality and dynamic vocabulary that is no longer under development. Consequently, where there are missing concepts needed for our work, there is no way to add them, thus necessitating the usage of other vocabularies as well.

our corpus. Entries may contain the orthographic word forms, phonetic forms (and variants), grammar, usage, sense, etymological information, links to relevant external lexical and knowledge resources, and related entries, as well as examples from the corpus. Additionally, a separate inflections dictionary document stores verbal paradigms.[14] In the following sections, we describe these features in more detail and demonstrate their encoding in TEI.[15] The work is still ongoing and the 567 entries in our dictionary (as of October 2018) represent only a small fraction of the total vocabulary in our corpus.

**Metadata and linking.** In addition to the typical features inherent in dictionaries listed above, through links declared in the header section (TEI Guidelines, The TEI Header) the TEI dictionary acts as a nexus of the linguistic (lexical feature inventories) and other referenced resources (e.g., personographic, bibliographic).

TEI allows numerous ways of linking to important information that may need to be referenced throughout a dictionary, and we make use of several approaches based on the type of reference and the data itself. In the following section we describe several such aspects of the dictionary and how they are relevant within the context of the language documentation. Figure 1 provides an overview of the linked and embedded resources in the Mixtepec-Mixtec dictionary at the heart of this project. Lexical features and terminology inventory. As mentioned above, the project adheres to standards as much as possible in all aspects of the work. Our inventory of lexical terminology is kept in a separate document containing TEI feature structures[16] that are used to tag the corpus. Figure 2 shows the declaration of the link to the document contained in the <sourceDesc> of the header in the dictionary (left) and a sample of two particular sets of features (trajector and landmark)[17] from the document it links to.

[14] MIX nouns and, in particular phrasal contexts, certain adverbs may be inflected for possession or other morpho-semantic features, and thus noun paradigms may also appear in the separate document.

[15] Note that the dictionary is still undergoing editing and at the time of submission the formatting discussed herein is not yet universally applied.

[16] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/FS.html

[17] Currently there exist no registered entries for these concepts in any public terminological repository, and they are among the list of proposals to be submitted for inclusion in the future.

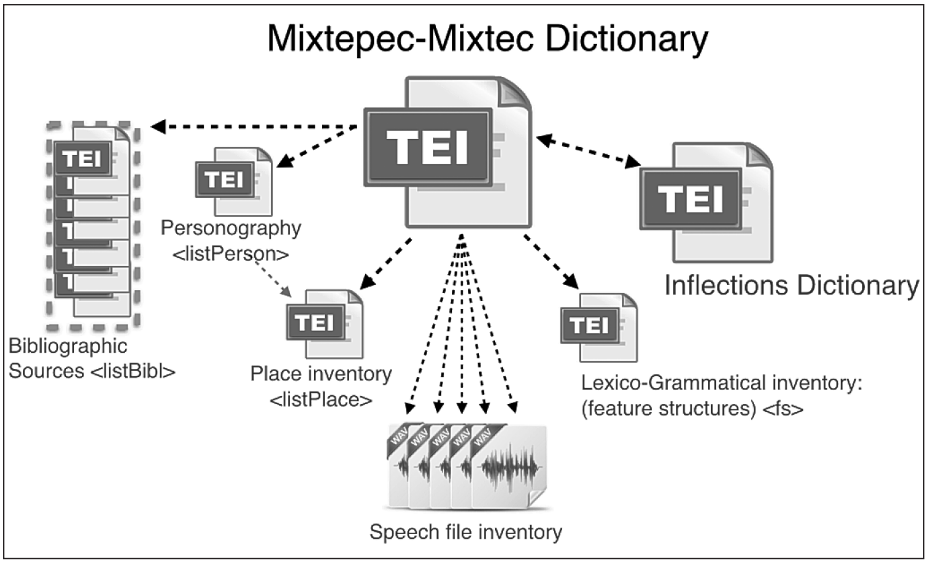


***FIGURE 1*** Diagram of Linked and Embedded Resources

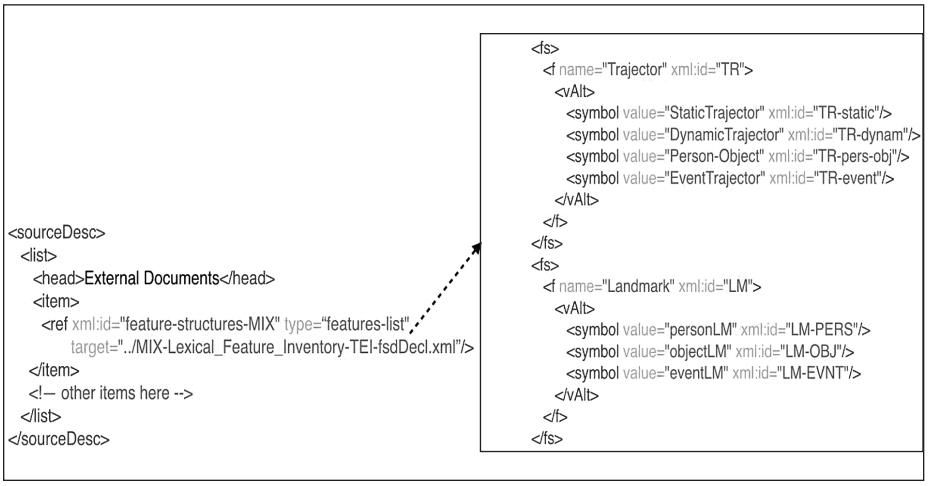


***FIGURE 2*** Feature Structures Declared in TEI Header and Their Content in Separate Document Bibliographic sources. External data such as documents by SIL make up a significant portion of our data. Within our dictionary we often need to point to these sources to attribute provenance of an example. To enable this we again declare these sources in the <sourceDesc> of the header. The pathway to the TEI file is declared in @xml:base as shown in Figure 3.

```
<listBibl xml:id="SIL-MEX">
  <head>SIL Mexico Publications</head>
  <bibl xml:id="bibl.L093" xml:base="SIL_docs/L093/L093-tok.xml">
    <title>Kunka'vi hora ka</title>
    <editor>Beckmann, Gisela</editor>(translator); <editor>Nieves, María
    M.</editor>(translator). <date>2007</date>. <edition>(2nd
      ed.)</edition>. <publisher>Instituto Lingüístico de Verano, A.C.</publisher>
    <pubPlace>Tlalpan, D.F., México</pubPlace>
  </bibl>
  <bibl xml:id="bibl.L094" xml:base="SIL_docs/L094/L094-tok.xml">
    <title>Kunchau hora ka</title>
    <editor>Beckmann, Gisela</editor>; <editor>Gómez Hernández, María</editor>.
      <date>2007</date>. <edition>(2nd ed.)</edition>. <publisher>Instituto
      Lingüístico de Verano, A.C.</publisher>
    <pubPlace>Tlalpan, D.F., México</pubPlace>
  </bibl>
  <!-- other bibl entries here-->
</listBibl>
```

***FIGURE 3*** <sourceDesc> and <listBibl>

**Personography.** Additionally, as shown in Figure 4, in the header (within <particDesc> embedded in <profileDesc>), we list each person (speakers, editors, and researchers) who may be referred to directly in the dictionary. This list also links to the external TEI personography document "MIX-People.xml" containing detailed information about the participants, the path to which is declared in @xml:base on the element <listPerson>.

```
<profileDesc>
  <particDesc>
    <listPerson xml:base="MIX-People.xml">
      <person xml:id="TS" role="speaker" corresp="MIX-People.xml#TS">
        <name>Juan "Tisu'ma" Salazar</name>
      </person>
      <person xml:id="JS" role="speaker" corresp="MIX-People.xml#JS">
        <name>Geremaia Salazar</name>
      </person>
      <person xml:id="JB" role="editor, researcher" corresp="MIX-People.xml#JB">
        <name>Jack Bowers</name>
      </person>
      <!-- more people here -->
    </listPerson>
  </particDesc>
</profileDesc>
```

***FIGURE 4*** <particDesc> and <listPerson> for Persons

Linking to related documents in dictionary. In certain entries we link to an external supplementary inflections dictionary containing inflectional paradigms. Within the TEI data structure shown in Figure 5, this is enabled by using the <prefixDef>[18] element in the header in which a prefix is declared and serves as a shortcut for a specific path within the inflections dictionary. This enables us to point to entries for a particular paradigm entry in the inflections dictionary, thus linking a lemma with its infections. In Figure 5, the value of @matchPattern is a template for such pointers with the regular expression ([a-zA-Z0-9]+), which is replaced by the specific text of an entry. At the end of the value of @ replacementPattern, #$1 means that any pointer with the prefix "paradigm" should point to the document "../MIX-Paradigms.xml" with the value of the first regular expression: ([a-zA-Z0-9]+).

```
<listPrefixDef>
   <prefixDef ident="paradigm" matchPattern="paradigm([a-zA-Z0-9]+)-V-paradigm-MIX"
              replacementPattern="../MIX-Paradigms.xml#$"/>
 </listPrefixDef>
```

***FIGURE 5*** Declaration of the <prefixDef> Pattern for Linking between Documents

Thus, within the dictionary, the inflections dictionary can be referenced by prefixing "paradigm:" within the string of pointer value. The pointer in Figure 6 links to the entry containing paradigms for the verb *kusu* 'sleep' (see Figures 12 and 13 for the context in which it is used).

```
<ptr type="inflectionParadigm" target="paradigm:sleep-V-paradigm-MIX"/>
```

***FIGURE 6*** Using Prefix Definition to Reference Verbal Paradigm Entry in Inflections Dictionary

**Other types of links.** Additionally, we have other resources such as sound files and some videos for which records, URLs, and metadata are stored in external files in a separate directory. Because of their large and ever-growing numbers, these are not declared in the header but can nonetheless be referenced and linked to within the dictionary using

[18] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-prefixDef.html

the @source attribute which can be placed in a number of different TEI elements.

## STRUCTURE AND CONTENT OF ENTRIES

**Forms and grammar.** The lemma of a MIX form is given in the orthographic form and, if attested to a high enough degree of certainty, in the phonetic form (IPA) as well. Each entry has a <gramGrp>, with the minimal information of part of speech and other features where applicable. The element containing the form always includes the @ xml:lang attribute, the value of which is the ISO 639 language tag.[19] If an abbreviated value is used, a @norm attribute with the full form of the feature is given in order to align with terminological standards. A typical example is shown in Figure 7.

```
<form type="lemma">
  <orth xml:lang=" mix">katsi</orth>
  <pron xml:lang="mix" notation="ipa">ká͡tsi</pron>
</form>
<gramGrp>
  <pos>verb</pos>
  <gram type="transitivity" norm="transitive">trans</gram>
</gramGrp>
```

***FIGURE 7*** Typical <form> and <gramGrp> Section of an Entry

**Variation, uncertain, and conflicting forms.** As this is a language documentation project, and the language is under-resourced both in its use as a literary language and its linguistic description, we are interested in recording variation and areas of uncertainty of all kinds.

**Orthographic variation**. Given that the MIX orthography is still under development and significant changes have been made over the last ten years, there are many lexical items in earlier documents with spellings that have since been changed. In these cases, both the old and up-to-date forms are represented. In the earlier publications (encoded herein

[19] The values of English and Spanish used are ISO 639-2; that of Mixtepec-Mixtec is from ISO 639-3 as that is the only option.

as the variant form), lexical tone was not represented in the orthography; however, this created a large number of homographs (which in some cases were of the same part of speech or even within the same domain). These needed to be distinguished.

The example in Figure 8 shows the updated *chuún* and antiquated *chuun* forms of the word meaning 'chicken' [ʧùú́], which is a tone-based minimal pair with the word meaning 'money' [ʧùù], the latter retaining the original spelling while the former adds the accent above the second vowel. The old form is labeled with <form type="variant"> and the element <orth> with attribute-value pair of @notation="plain" to indicate that the orthography does not represent tone.

```
<form type="lemma">
  <orth xml:lang="mix">chuún</orth>
  <form type="variant">
    <orth notation="plain" xml:lang="mix">chuun</orth>
  </form>
  …
</form>
```

***FIGURE 8*** Entry with Alternative Spelling without Tonal Diacritics

Additionally, given that the orthographic standard being developed has not been published,[20] those who write in the language often do not use the working spelling conventions, and thus we are faced with inte-grating all variants into our common system. The example in Figure 9 shows the encoding of a variant spelling of the lexical item meaning 'water' which was observed in a public service publication by the Mexican government. In this orthography, the voiceless alveo-palatal affricate is rep-resented as *ty* instead of the standard *ch*, and the long word-final vowel is represented only as a single *i*. The document that is the source of the spelling variant is provided as the value of the @source attribute, which is declared in the bibliography within <listBibl> in the header (see "Linking to related documents" under "Metadata and linking" in TEI DICTIONARY section).

[20] The latest known update to the orthography was obtained via personal communication with Mille Nieves of SIL Mexico in June 2017; it is upon this version that all editorial practice is based with regard to spelling normalization.

```
<form type="lemma">
  <orth xml:lang="mix">chikuii</orth>
  <form type="variant">
    <orth source="#infografica-308-inundaciones" xml:lang="mix">tykui</orth>
  </form>
  ….
</form>
```

***FIGURE 9*** Variant Orthography from MIX Language Publication

**Phonetic variation.** In our data there are certain lexical items for which pronunciation variants are observed frequently enough that alternate pronunciations are included in the dictionary entry. In the example shown in Figure 10, the primary pronunciation[21] (with the onset voiceless alveo-palatal fricative) is placed as a direct child of <form>, and the variant (with the onset voiced alveo-palatal fricative) is embedded within a separate <form>, also labeled @type= "variant".

```
<form type="lemma">
  <orth xml:lang="mix">chikuii</orth>
  …
  <pron notation="ipa" xml:lang="mix">ʧikʷìì</pron>
  <form type="variant">
    <pron notation="ipa" xml:lang="mix">ʤikʷìì</pron>
  </form>
  ….
</form>
```

***FIGURE 10*** Encoding of Phonetic Variants

**Integrating data from external sources.** Despite there being only a small body of linguistic literature about the language, there are cases where examples of transcribed vocabulary found in such sources are of interest and are thus integrated into the dictionary. Some instances may be the first or only attestation of the word or may diverge in some way from our own characterizations of the item. Additionally, there may be divergence in the transcription conventions used to represent the content.

[21] The primary pronunciation, where present, is determined by weighing the factors of observation frequency and knowledge of the language's phonology.

One such example involves the tone of *iin* 'nine', where data from our project differs from that of two previous papers. In twenty-four of the twenty-five tokens in our data (from three speakers), the F0 (pitch) analysis shows that the tonal contour is falling, while the twenty-fifth, from a fourth speaker, is ambiguous. Thus, based on our own observations, we characterize the tone as *falling*. However, both Pike and Ibach (1978) and Paster and Beam de Azcona (2004a)[22] characterize the tone simply as low.[23] Figure 11 shows the encoding of this in the dictionary. Since the statistics support our description we use the certainty attribute with the value "high" on this form.

```
<form type="lemma">
  <orth xml:lang="mix">iin</orth>
  <pron notation="ipa" xml:lang="mix" cert="high">ĩĩ↘</pron>
  <pron notation="ipa" xml:lang="mix" source="#bibl.pike-ibach-1978" orig="j³j³">ĩĩ˩˥</pron>
  <pron notation="ipa" xml:lang="mix" source="#paster-azcona-2004a" orig="íi̲">ĩĩ˩˥</pron>
  ....
</form>
```

***FIGURE 11*** Entry with Conflicting Phonological Descriptions in External Sources

Another noteworthy observation in this example is the treatment of transcription notation. Unfortunately, nearly none of the past studies of MIX phonology used IPA notation in their transcriptions. Fortunately, TEI has the ability both to keep the original forms from the sources in

[22] The speaker consultant for the Paster and Beam de Azcona studies is the same individual as the current project's primary speaker (Juan "Tisu'ma" Salazar); we are thus confident that this is an issue of diverging phonological descriptions rather than a difference in pronunciation. Given the extensive body of work we have carried out and the fact that our transcriptions are based on recordings, we can dispute this question with confidence.

[23] There are still questions as to the degree of tonal contrasts that remain to be answered via further systematic study of the phonology. In particular it is not yet fully clear whether there are minimal phonological distinctions based on each degree of tonal contour; that is, *mid-low* vs *high-low*, or even whether there is a contrast between *falling* and *low* tones. For this reason in our IPA transcription of what we characterize as phonologically *falling*, we currently use the global fall arrow ↘ as a temporary placeholder instead of the combining diacritic ˕̃ or the combined tone letters ˥˩ or ˧˩.

@orig and to normalize the notation to IPA in the element values for compatibility.

**Inflection and paradigms.** As mentioned above, a separate inflections dictionary contains full inflectional paradigms to which entries can link using the TEI <prefixDef> strategy described earlier. This is done in the primary dictionary[24] with the <ptr> element as shown in Figure 12.

```
<form type="lemma">
  <orth xml:lang="mix">kusu</orth>
  <ptr type="inflectionParadigm" target="paradigm:sleep-V-paradigm-MIX"/>
</form>
<gramGrp>
  <pos>verb</pos>
</gramGrp>
```

***FIGURE 12*** Verbal Lemma Entry with Link Using <prefixDef> Pattern to Link to Paradigms

The pattern of @xml:id values in the paradigm entries is (verb name)-V-paradigm-MIX (e.g., sleep-V-paradigm-MIX). Thus, the pointer in Figure 13 identifies a specific entry in the MIX-Paradigms.xml.

```
<entry xml:lang="mix" xml:id="sleep-V-paradigm-MIX">
```

***FIGURE 13*** Example of Paradigm Entry Target Referred to in Primary Dictionary

In MIX, inflections can occur on verbs, nouns (for possession), and adverbs (in certain phrasal contexts).[25] Within the form section, full paradigms are represented as embedded blocks of inflected forms in accordance with the recommendations of TEI Lex0 (Bański et al. 2017). Each paradigm is encoded as a sibling of the lemma in <form type=”paradigm”> and the primary common feature (tense or mood) is

[24] “Primary dictionary” refers to the main Mixtepec–Mixtec dictionary as distinct from the inflections dictionary.

[25] To our current knowledge, only the adverb *ncho'a* can be inflected (for person) and only in certain lexicalized phases in which it occurs with a copular adjective or verb phrase.

labeled as the value of @subtype, and tense/aspect/voice are encoded in <gramGrp>. In Figure 14, the first two forms of the paradigm for the present/incompletive forms of the verb *kusu 'sleep'* are shown.

```
<form type="paradigm" subtype="present">
  <gramGrp>
    <tns norm="present">pres</tns>
    <gram type="aspect" norm="incompletive">incmpl</gram>
    <mood norm="indicative">indic</mood>
  </gramGrp>
  <form type="inflected">
    <orth xml:lang="mix">kíxi yu</orth>
    <pron xml:lang="mix" notation="ipa">kíʃi jù</pron>
    <gramGrp>
      <per>1</per>
      <number norm="singular">sg</number>
    </gramGrp>
  </form>
  <form type="inflected">
    <orth xml:lang="mix">kíxu</orth>
    <gramGrp>
      <per>2</per>
      <number norm="singular">sg</number>
      <gram type="register" norm="informal">inf</gram>
    </gramGrp>
  </form>
 <!-- other forms here -->
</form>
```

***FIGURE 14*** Partial Verbal Paradigm for MIX Verb *kusu*

Note that there is a <gramGrp> as a direct child of <form type="paradigm">, and this contains the grammatical information common to all the inflected forms in the paradigm and inherited via the inheritance principle (Ide et al. 2000).

## SENSE

The <sense> section contains information pertaining to meaning, including definitions, translations, examples of usage in context, domain classification, and a number of other data fields pertaining to semantic relations. An entry can have any number of senses.

Links to external knowledge sources: dbpedia. In order to enrich the content of our dictionary, for each concept (where available) we insert a link to an entry in dbpedia (Auer et al. 2007) or other open knowledge resources in the @corresp within the sense element, as shown in Figure 15.


***FIGURE 15*** Linking Sense to dbpedia

This is done with several benefits in mind. One is that they provide a link between a structured body of human knowledge and the Mixtepec-Mixtec language. Currently there are no Mixtec language wiki resources, and these links to dbpedia could provide a template upon which a MIX version of wiki-type entries could be based. Additionally, the multilingual definitions of the concepts found in the entries could serve as a systematic reference point upon which to base MIX definitions of the senses, which (as discussed below) are currently available for a small number of entries. Finally (with the inclusion of @xml:id) they enable the compatibility of the data to linked data formats such as OntoLex-Lemon (McCrae et al. 2017).

**Translations.** The most basic facet of the sense section is the multilingual translations into English and Spanish. Translations of lemmas are placed <cit type="translation"> within the <form><orth> element block. If in the translation language the Mixtec item has more than one specific translation, the others are listed in separate <cit> elements. Where possible we may also include links to digital dictionary-external resources for the given translation target languages, as in Figure 16.[26]

[26] The use of Wiktionary as a source of enhanced translations is a work in progress and is not yet systematically applied to each instance. In the future we may seek an automated means of gathering and adding this data.

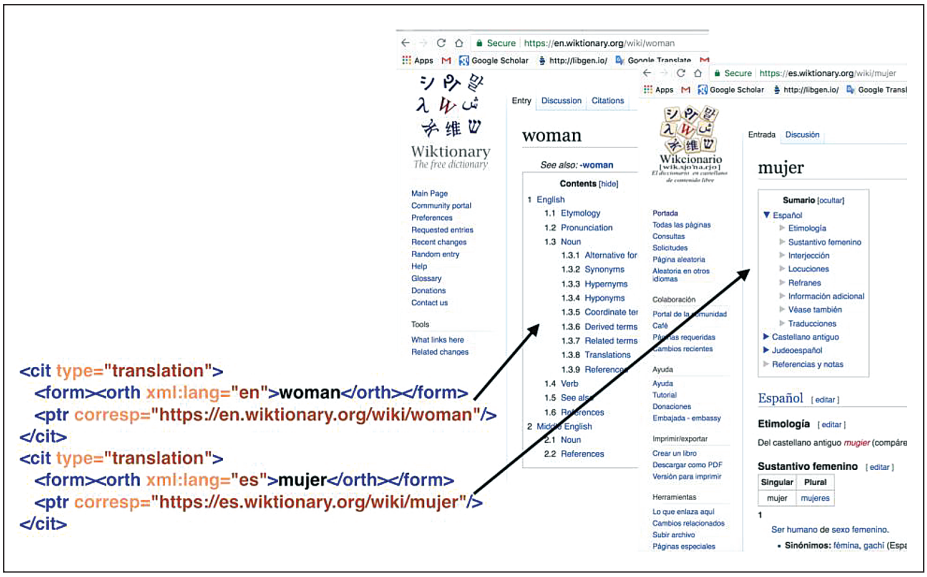


***FIGURE 16*** Linking Translations to Wiktionary Entries

**Examples.** Any number of examples of the usage of an item in the context of the source data can be included within sense; these are also encoded as <cit> with the @type="example" and the wrapper <quote>. The form in question is wrapped within an orthographic reference element <oRef>. A pointer to the source of the example is included as the value of @target in the pointer <ptr/> element, as in Figure 17.

```
<cit type="example">
  <quote xml:lang="mix">¿A tsinu kue <oRef>tsa'a</oRef> kiti?</quote>
  <ptr target="SIL_docs/L097/L097-tok.xml"/>
</cit>
```

***FIGURE 17*** Lemma in Textual Context from Corpus with Link

**Definitions.** Entries can include definitions in Mixtec, Spanish, and English. A major goal is to have definitions for senses in all three languages to allow for the creation of monolingual, bilingual, and trilingual dictionaries. At the present stage, however, most entries do not have a Mixtec-language definition. In such cases we create a simple Spanish or English definition that can be used as a template for a future Mixtec one. In those cases we include separate <def> elements for each language. Figure 18 shows a sub-sense of *nuu* 'face', which is used to express the sense 'front of (something)'.

```
<def xml:lang="en">The front of (sth).</def>
<def xml:lang="es">La parte delantera de (algo).</def>
<def xml:lang="mix"/>
```

***FIGURE 18*** Multilingual Definitions with Empty Element for Mixtec to be Added

**Images.** In certain entries (often ones that correspond with certain theoretical interests pertaining to metaphor-driven and metonymy-driven sense change), we may include images showing the concept denoted in the sense. In TEI this is done with <graphic @url>, within which the <desc> element describes the content of the image. As in <def>, we include English and Spanish along with an empty tag for a future Mixtec description to be added. These images could be used for a pictographic or multimedia learning resource (e.g., a children's dictionary) or as examples for our own presentations. Figure 19 shows a visualization of the given sense of the word for 'face', which in this sense means 'front of' something.

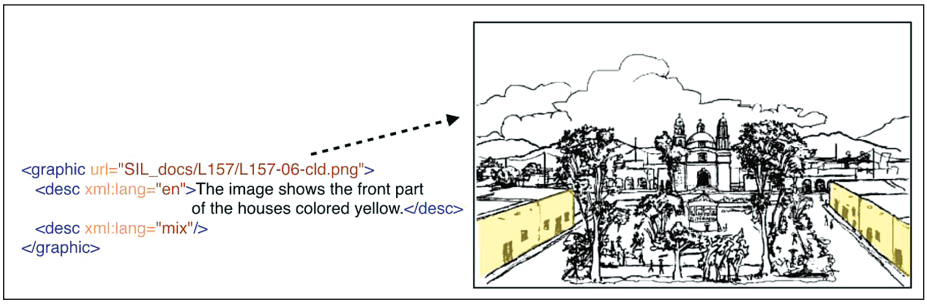


***FIGURE 19*** FIGURE 19 Image Modified to Illustrate an Extended Sense of the Word for 'face'

**Semantics and cultural issues in language documentation.** Especially in a language documentation project, it is important and necessary to include other notes on various specifics of an entry. An example is the lexical item *sa'an ntavi*, one of two terms referring to the Mixtepec-Mixtec language itself and whose components translate literally as 'poor language'. Our project's native speaker consultant (understandably) finds this term offensive and derogatory and wanted it marked as dispreferred in the dictionary and the issue to be recorded in prose. This

is achieved with a combination of the TEI <note> and <usg> elements, with the @type value of "attitude" and the @resp specifying the initials of those responsible for recording this information, as shown in Figure 20. The initials are the @xml:id value of the individuals and are declared in the header (as discussed in "Personography" under "Metadata and linking" in TEI DICTIONARY section).

```
<usg type="attitude" resp="#TS">dispreferred</usg>
<xr type="synonymOf"><ref xml:lang="mix" target="#MIX-language-rain">sa'an savi</ref></xr>
 <note resp="#TS #JB">This term which translates as "poor language" is dispreferred by speakers
consulted as it is derogatory. This is so particularly in contrast to the term for the Spanish language
<oRef>sa'an xchila</oRef> which translates as "fancy language".</note>
```

***FIGURE 20*** Components of Entry Detailing Dispreferred Status of Lemma with Pointer to Synonym[27]

**Semantic relations and domain.** In addition to sense, translations, and definitions, our dictionary includes information on semantic relations and domain. While the former is commonly made use of in structuralist linguistic approaches and computational linguistics such as WordNet (Miller 1995; Fellbaum 2005, 2010), the latter is typical of theoretical approaches based in cognitive linguistics (Langacker 1987; Clausner and Croft 1999).

While within theoretically these features are a mixture of structuralist and encyclopedic models of semantics (Geerarts 2010), for the purposes of the project, including these features in the annotation brings significant benefits. From the point view of potential Mixtec users of this resource, these features can be harnessed to facilitate collection and generation of focused sets of vocabulary to be used for the creation of further, more focused resources such as children's books and thesauri. Below we describe the content and implementation of these features in our dataset.

Semantic relations in the dictionary are encoded within specific senses of an entry within the external relation element <xr>. The given type is encoded in @type with an embedded <ref> that takes the @xml:lang (as we are generally providing both Mixtec and

[27] See following section (Semantic relations) for explanation of the use of <xr type="synomymOf">.

English versions, with English being the metalanguage for computational purposes). Where they point to other entries within the dictionary the @target attribute is used on <ref>. In the dictionary we tag only the members/subclasses but not the top nodes; thus, in the entry for 'fruit' we do not have the semantic relation 'hypernym' for every specific fruit species. Instead, this collection can be inferred and built up from the body of items tagged "hyponym" of 'fruit'.

Hyponymy is realized as <xr type="hyponymOf">. This category is extremely useful for generating taxonomical vocabulary lists. For the semantic relations hyponymy and meronymy, an additional <ref type="sense"> is included with the @corresp, the value of which is the same as occurs on that item's sense element. Thus, for the entry for 'peach' or other type of fruit, the <ref type="sense"> contains the same dbpedia URL as does the <sense @corresp> entry for *kui'i* 'fruit' itself, as shown in Figure 21.

```
<xr type="hyponymOf">
  <ref target="#fruit-MIX" xml:lang="en">fruit</ref>
  <ref target="#fruit-MIX" xml:lang="mix">kui'i</ref>
  <ref type="sense" corresp="http://dbpedia.org/resource/Fruit"/>
</xr>
```

***FIGURE 21*** <xr> for Hyponym Relations as in Entries for Types of Fruit

Meronymy is realized as <xr type="meronymOf"> [28]. As discussed by Geeraerts (2010), meronymy and hypernymy are central to the realiza-tion and analysis of metonymy. Synonymy and antonymy are encoded as <xr type="synonymOf"> and <xr type="antonymOf">. There are limits, however, to semantic relations both functionally and theoretically, as not all relevant semantic correlations in vocabulary or (more importantly) human knowledge can be defined or linked together in terms of hierarchical or pure opposition or identical senses. In order to fill some of that gap, we also use the concept of semantic domain.

[28] While meronymy can be and in a number of theoretical sources is subtyped according to different conceptual paradigms, there are theoretical conflicts (Geeraerts 2010) as to the soundness of these distinctions; until further research and evaluation of this question can be carried out, then, we will not assign such sub-typologies.

In addition to semantic relations, which in lexicography are more immediately useful in computational applications, where applicable we assign semantic domain (Langacker 1987; Clausner and Croft 1999) to the sense of certain entries, a fairly common practice in compiling dictionaries. In lexicographic practice, however, the use of domains in a dictionary is often limited to technical subject classes (e.g., medicine, zoology, literature). Domains are fundamental cognitive concepts according to which humans organize, understand, and represent experience and knowledge of the world (Langacker 1987; Clausner and Croft 1999), and this is a particularly enriching perspective in approaching language documentation.

In cases of polysemy, semantic domain is often a key distinction between the various senses. In Figure 22 we show the senses in the entry for *kani* 'long' (domain of SPACE), which can be used in the sense of the domain TIME. In TEI, domain is encoded as <usg type="domain">.[29]

```
<sense n="1" xml:id="long-space">
  <usg type="domain">Space</usg>
  <cit type="translation">
    <form><orth xml:lang="en">long</orth></form>
  </cit>
  <cit type="translation">
    <form><orth xml:lang="es">lungo</orth></form>
  </cit>
  <sense n="2" xml:id="long-time">
    <gramGrp>
      <pos>adv</pos>
    </gramGrp>
    <usg type="domain">Time</usg>
    <cit type="example">
      <quote xml:lang="mix" resp="#TS"><oRef>Kani</oRef> nchu'a ntsi ra.</quote>
      <cit type="translation">
        <quote xml:lang="en">He lived a long time.</quote>
      </cit>
      <!— spanish translation here -->
    </cit>
<!— <etym> here -->
</sense>
```

***FIGURE 22*** Two Senses in the Entry for *kani* 'long'

[29] Where available, like <sense> and <xr>, domain may also include URLs from external ontologies or sources such as dbpedia.

The inclusion of semantic domain potentially enables an alternate system of organization of a dictionary from the typical alphabetical ordering, or a derived domain-specific dictionary, and it can provide assistance with both manual and automatic word sense disambiguation (WSD).[30] Finally, domains enable us to encode and provide more dynamic analyses of sense-based etymological processes in keeping with cognitive linguistic theory. This latter is particularly important to the description of Mixtecan languages, as discussed in the following section.

## ETYMOLOGY

In addition to the general documentation of the language, this project's dictionary is being created as a structured database of etymological information. In our data we have observed and encoded the full array of etymological processes, including borrowing (mostly from Spanish, some from Nahuatl), inheritance (from a posited Proto-Mixtecan language inferred by comparing cognates), and form changes: compounding, derivation, phonological change; various types of sense change such as metaphor, metonymy, and grammaticalization, as well as numerous instances of combinations of these processes. Bowers and Romary (2018) discuss such phenomena and their encoding in TEI.

**Sense-related etymologies.** As mentioned, a major point of emphasis in this project is the semantics, specifically the strategies of lexical innovation, particularly from the perspective of cognitive linguistics. There exists a significant body of literature discussing the evidence of metaphor and metonymy in lexical innovation in related varieties of Mixtecan (Hollenbach 1995; Brugman and Macaulay 1986; Langacker 2002); the dataset for MIX provides ample content that enriches such linguistic discussions (Bowers 2016 and forthcoming).

Figure 23 shows the etymology for MIX *kani* 'long' in the sense of the domain of TIME (discussed in the previous section). The conventions shown in the encoding of such phenomena in TEI have been derived from the recommendations set forth in Bowers and Romary (2016) and

[30] Word sense disambiguation is particularly important given that the MIX orthography represents tone only on a small percentage of words.

are in accordance with those in the etymology section of TEI-Lex0[31] (Bowers and al. 2018).

```
<sense n="1" xml:id="long-space">
  <usg type="domain">Space</usg>
  <cit type="translation">
    <form><orth xml:lang="en">long</orth></form>
  </cit>
  <cit type="translation">
    <form><orth xml:lang="es">lungo</orth></form>
  </cit>
  <sense n="2" xml:id="long-time">
    <gramGrp>
      <pos>adv</pos>
    </gramGrp>
    <usg type="domain">Time</usg>
    <cit type="example">
      <quote xml:lang="mix" resp="#TS"><oRef>Kani</oRef> nchu'a ntsi ra.</quote>
      <cit type="translation">
        <quote xml:lang="en">He lived a long time.</quote>
      </cit>
    </cit>
    <etym type="metaphor" cert="high">
      <seg type="desc">Active zone of source profile (aka ontological knowledge/
impetus) motivating the metaphor is QUANTITY.  The domain mapping directionality of the
sense change is: QUANTITY of SPACE (SIZE or DISTANCE) → QUANTITY of TIME. The
domain shift is thus: SPACE → TIME. This directionality is predictable as it follows the pattern of:
CONCRETE → ABSTRACT; and of which, the formost is SPACE →  TIME.</seg>
      <cit type="etymon" corresp="#long-space">
        <usg type="domain">Space</usg>
      </cit>
    </etym>
  </sense>
</sense>
```

***FIGURE 23*** Metaphorical Sense Change 'long' (SPACE > TIME)

Despite having no written evidence of this lexical item in earlier stages of the language, we are able confidently to assert the directionality of this relationship between these senses, as the metaphorical process of SPACE > TIME is a predictable mapping that follows the general pattern of utilizing concrete conceptual structures to describe and understand abstract concepts (Kövecses 2010; Gentner et al. 2002; Boroditsky 2000).

Herein the sense of 'long' (TIME) is embedded within the first—spatial—sense, which in this dictionary is done where one sense is clearly

[31] The TEI-Lex0 project is taking place within the DARIAH lexical data working group.

derived from another. When there is one or more embedded <sense> elements, the respective etymologies within should be considered sequential, stemming from the highest sense. In our example, they are also numbered using @n. On the etymology element <etym>, we use the @type to classify the etymological processes. If there is more than one sub-process involved, we can use embedded <etym> element structures to represent them.

We provide a prose description of our analysis of the given process in the <seg type="desc"> element. Given that it is a polysemy and is the same form as the source sense, the etymon <cit type="etymon"> does not have a form in this case. The @corresp attribute points to the source of the sense change that is the first sense. In addition to the @type="metaphor", the data structure contains the key information for that process in the <usg type="domain">, which are in both senses, and copied within the <cit type="etymon">. Together with the embedding of senses and etymology, the contrast in the domain values from the first sense to the second provides a set of structured data that can be computationally searched and summarized.

## FUTURE ENDEAVORS

This project is the subject of an ongoing PhD thesis by the first author, and, moving forward, funding will be sought to further work towards several major goals, including:

- seeking partnerships to bring in long-term Mixtec contributors and editors, with the goals of expanding the size of the vocabulary and adding MIX iterations of key core contents of the dictionary (particularly the definitions and descriptions of the etymology);
- inviting additional linguists to make use of the extensive data collected from specific linguistics subfields;
- integrating other relevant Mixtecan resources into the dictionary and corpus using OCR with GROBID dictionaries (Khemakhem et al. 2017), namely the Colonial Mixtec vocabulary collection by fray Francisco de Alvarado from 1593 (Jansen and Perez 2009).

## CONCLUSION

In this paper we have described the structure and contents of a multilingual TEI dictionary, which is part of an ongoing language documentation project for the Mixtepec-Mixtec language, an indigenous language of Mexico. We have shown how we are using the TEI dictionary to accommodate a number of key linguistic and metadata-related needs with a particular focus on issues associated with the documentation of an under-resourced language. In addition, we have described how we are integrating the MIX lexical resources with structured external knowledge sources such as dbpedia, and how within this digital dictionary we are working towards building a database of semantic and lexical relations, as well as etymological data. Finally, in using TEI for a language documentation project, we hope that our work may provide a useful reference for prospective projects whose researchers might be seeking guidance for other language and project-specific issues.

As a final note, Figure 24 shows a relatively complete dictionary entry in the print view for *antivi* 'sky', with various elements present in the entry. The print rendition is done using CSS (Cascading Style Sheets). Note that not all of the contents present in TEI are rendered in the print view: certain features such as the semantic relations are intended for computational purposes and are not necessarily relevant to the dictionary's primary target audience for the print view, principally members of the Mixtec community. The underlined contents signify they are linked to external or internal locations that may be source files or a URL.

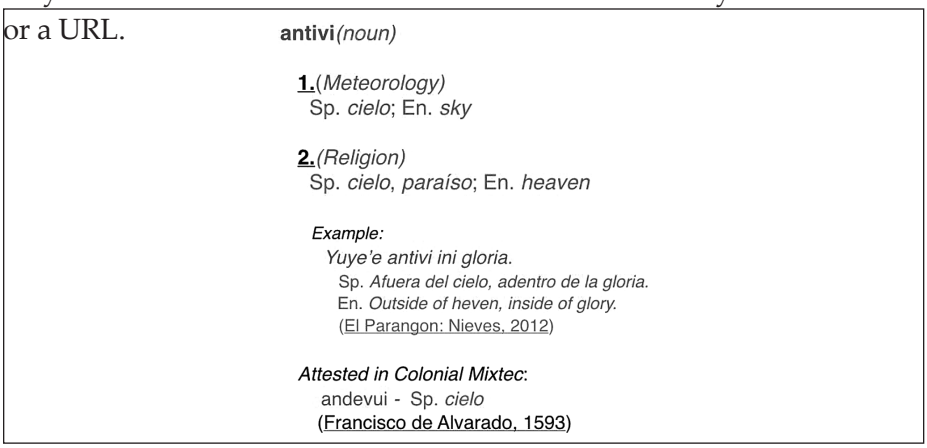


***FIGURE 24*** Example of Print View of Entry *antivi* as Rendered Using CSS

## REFERENCES

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, edited by Karl Aberer et al. Lecture Notes in Computer Science, Vol. 4825. Berlin and Heidelberg: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-76298-0_52

Austin, Peter K. 2006. Data and language documentation. In *Essentials of Language Documentation*, edited by Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 87–112. Berlin: Mouton de Gruyter.

Bański, Piotr, Jack Bowers, and Tomaž Erjavec. 2017. TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. In *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference*, edited by Iztok Kosem et al., 485–94. Available at https://elex.link/elex2017/proceedings-download/

Boersma, Paul, and David Weenink. 2017. Praat, a system for doing phonetics by computer (Version 6.0.28). Retrieved from http://www.praat.org/

Boroditsky, Lera. 2000. Metaphoric structuring: Understanding time through spatial metaphors. *Cognition* 75.1: 1–28.

Bowers, Jack. 2016. A cognitive analysis of Mixtepec-Mixtec body part terms. In *La grammatical de las expressiones de partes del cuerpo*. Lima, Peru: PUCP.

——. Forthcoming. Pathways and patterns of metaphor, metonymy in Mixtepec-Mixtec body-part terms.

Bowers, Jack, Axel Harold, and Laurent Romary. 2018. TEI-Lex0 Etym – towards terse recommendations for the encoding of etymological information. Presented at the *TEI Conference and Members' Meeting*. Tokyo, Japan.

Bowers, Jack, and Laurent Romary. 2016. Deep encoding of etymological information in TEI. *Journal of the Text Encoding Initiative* (Issue 10). https://doi.org/10.4000/jtei.1643

——. 2018. Encoding Mixtepec-Mixtec etymology in TEI. Presented at the *TEI Conference and Members' Meeting*. Tokyo, Japan.

Brugman, Claudia, and Monica Macaulay. 1986. Interacting semantic systems: Mixtec expressions of location. In *Annual Meeting of the Berkeley Linguistics Society* 12: 315–27.

Clausner, Timothy C., and William Croft. 1999. Domains and image schemas. *Cognitive Linguistics* 10: 1–32.

Czaykowska-Higgins, Ewa , Martin D. Holmes, and Sarah M. Kell. 2014. Using TEI for an endangered language lexical resource: The Nxaʔamxcín Database-Dictionary Project. *Language Documentation & Conservation 8: 1–37*.

Farrar, Scott, and Terry Langendoen. 2003. A linguistic ontology for the semantic web. *Glot International* 7.3: 97–100.

Fellbaum, Christiane. 2005. WordNet and wordnets. In *Encyclopedia of Language and Linguistics*, 2nd edn., edited by K. Brown, 665–70. Oxford: Elsevier.

——. 2010. WordNet. In *Theory and Applications of Ontology: Computer Applications*, edited by Roberto Poli, Michael Healy, and Achilles Kameas, 231–43. Dordrecht: Springer.

Geeraerts, Dirk. 2010. *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Gentner, Dedre, Mutsumi Imai, and Lera Boroditsky. 2002. As time goes by: Evidence for two systems in processing space→time metaphors. *Language and Cognitive Processes* 17.5: 537–65.

Hollenbach, Barbara E. 1995. Semantic and syntactic extensions of body-part terms in Mextecan: The case of 'face' and 'foot'. *International Journal of American Linguistics* 61.2: 168–90.

Ide, Nancy, Adam Kilgarriff, and Laurent Romary. 2000. A formal model of dictionary structure and content. In *Euralex 2000*, 113–26. Stuttgart, Germany. Retrieved from http://arxiv.org/abs/0707.3270

Ide, Nancy, and Laurent Romary. 2004. A registry of standard data categories for linguistic annotation. In *4th International Conference on Language Resources and Evaluation-LREC'04*, 135–38. Retrieved from http://hal.archives-ouvertes.fr/inria-00099858/

Khemakhem, Mohamed, Luca Foppiano, and Laurent Romary. 2017. Automatic extraction of TEI structures in digitized lexical resources using conditional random fields. In *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference*, edited by Iztok Kosem et al. Retrieved from https://hal.archives-ouvertes.fr/hal-01508868v2

Kövecses, Zoltán. 2010. *Metaphor: A Practical Introduction*, 2nd edn. Oxford: Oxford University Press.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Vol. 1. Stanford, CA: Stanford University Press.

——. 2002. A study in unified diversity: English and Mixtec locatives. In *Ethnosyntax: Explorations in Grammar and Culture*, edited by N. J. Enfield, 138–61. Oxford: Oxford University Press.

McCrae, John P., Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: Development and applications. In *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference*, edited by Iztok Kosem et al., 587–97. 19–21 September. Leiden, the Netherlands. Available at https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38.11: 39–41.

Paster, Mary. 2005. Tone rules in Yucanani Mixtepec Mixtec. Presented at the SSILA meeting, Oakland, CA. Available at http://pages.pomona.edu/~mp034747/SSILA2005.pdf

——. 2010. The role of homophony avoidance in morphology: A case study from Mixtec. *Proceedings from the 13th Annual Workshop on American Indigenous Languages. Santa Barbara Papers in Linguistics* 21: 29–39.

Paster, Mary, and Rosemary Beam de Azcona. 2004a. A phonological sketch of the Yucunany dialect of Mixtepec Mixtec. *Proceedings from the Seventh Workshop on American Indigenous Languages. Santa Barbara Papers in Linguistics* 15: 61–76.

——. 2004b. Aspects of tone in Yucunany Mixtepec Mixtec. Presented at the *Conference on Otomanguean and Oaxacan Languages*, 19–21 March. University of California at Berkeley.

Pike, Eunice V., and Thomas Ibach. 1978. The phonology of the Mixtepec dialect of Mixtec. In *Linguistic and Literary Studies in Honor of Archibald A. Hill, Volume 2: Descriptive Linguistics*, edited by Mohammed A. Jazayery, Edgar C. Polomé, and Werner Winter, 271–85. The Hague: Mouton.

Simons, Gary F., and Charles D. Fennig, eds. 2018. Mixtepec-Mixtec. In *Ethnologue: Languages of the World*, 21st edn. Dallas, Texas: SIL International. Available at https://www.ethnologue.com/language/mix

TEI Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.4.0. July 2018. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

TEI Consortium. "The TEI Header." TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.4.0. July 2018. TEI Consor-tium.http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html